Toward the Development of Next Generation Search Engine

Dwi Hendratmo Widyantoro*

Bandung Institute of Technology, Jalan Ganesa 10, Bandung 40132, Indonesia

Search engine has been the main source for finding information on the Internet. This document proposes a new search engine architecture that will help users locate relevant information currently overlooked by existing major search engines. The proposed search engine groups search results according to classes of information defined by a set of ontology relevant to user query, and provide users with interactive navigational interface to explore the search results. It also incorporates summarization extraction techniques to enhance user searching experiences. In order to make the system scalable, we propose a framework for collaborative, community-based ontology creation environment. The proposed search engine architecture is expected to be an embryo for next generation search engine.

1. Introduction

The importance of search engine is unquestionable. It has been the main tool for searching information, particularly for those who have been exposed by the Internet. The inventions of Google’s pageRank [3], hub & authority [11] and similar ranking algorithms in the last decade have given a boost to the search engine popularity. These algorithms share a common assumption in that the popularity of Websites (partly determined by the number of in-going links) is a good relevance measure. This assumption has been the main contributing factor, among others, to the success of major search engines such as Google [20], Yahoo! [21] and Microsoft Search [22].

Despite their success, we identify two problems in these search engines. First, the Website popularity assumption employed by major search engines presents an inherent weakness. Specifically, it will rank low to unpopular Websites even though their contents are highly relevant to search queries, for example, in newly developed Websites or personal Websites that never advertise their existence. It has been reported that PageRank is biased against new Web pages [2,6]. Second, a list of search results returned by major search engines, despite its simplicity, can hinder users from finding relevant information. Users are generally willing to examine only the first two pages of search results. A truly relevant hit ranked beyond this range will have a high chance undiscovered by users. The list of search results does not provide clue of which pages the users should look for, and hence discourages them to further exploration. In the mean time, the proliferation of information on the Internet will continuously increase the number of relevant information over time. The majority of them could be from unpopular, unheard-before Websites. The implication in the near future would be that the state-of-the-art technologies of current search engines still return relevant results, but the majority of relevant Websites (among of them could be the truly relevant ones) remain unexplored.

Motivated by the above problems, this paper proposes a search engine test bed that helps users quickly find relevant information, with respect to search queries, regardless of the popularity of information sources. We pursue two complementary research approaches to achieve this goal. The first approach is to group search results according to various criteria and provide navigational interface for search results exploration. This interaction style has been shown to be of helpful to users [4,9]. The second research approach is to develop appropriate technique for summarizing search results, allowing users to view accurate content of each result without having to visit the corresponding Website.

2. Related Works

There has been a number of existing works that are similar to ours. Vivisimo is such an example of commercial search engine [23]. Existing works mostly employ (incremental) clustering technique for grouping search results on-the-fly. Although scalable and functional, using a clustering method often produces confusing organization of search results category. Unlike prior works, we employ ontology to obtain better hierarchical organization of search results. We propose a novel search engine architecture that also indexes and maintains an ontology collection. Given a search query, the system retrieves both relevant ontology and document links. The retrieved ontology will then be used to hierarchically classify the actual search results. This paper also proposes a framework for defining and submitting new ontology to the system. Similar in spirit to that in the Wikipedia [26], the framework will allow others to contribute in the global-scale ontology development (particularly for the purpose of information classification).

This automated hierarchical classification distinguishes our proposed research with Open Directory search engine [25], which is manually labelled by human editors. If the topics of search results cannot be identified from ontology, standard clustering method will be employed to group the search results. Previous works organize information from various sources. DynaCat organized information using metadata extracted from UMLS thesaurus [16]. Allen used Dewey Decimal System [1]. Book content can also be adopted as the information structure to organize information such as in Marchionini [14] and SuperBook project [12]. Other researchers exploited Web Site links structure.

WebCutter, for example, presented Web Site map that was customized according to the user query [13]. AMIT system

* E-mail: dwi@if.itb.ac.id
grouped search results into a Web Structure [17]. Similar approach was also developed in Cha-Cha [5]. Unlike our approaches that focus on improving the presentation of search results, other researchers tackle the same problem by searching an unbiased Web ranking algorithm.

3. Overview of Search Engine Architecture

Fig. 1. depicts the search engine architecture developed in this research. It basically extends existing conventional search engines by incorporating ontology collection for automatic search results classification, clustering, and Website summarization for better presentation of search results.

This research employs LUCENE as the base search engine [8,24]. LUCENE is an open source software and has been widely used in many commercial applications. Its main functions in the system’s architecture are (1) to index Website and other documents collected from the Internet and (2) to provide search mechanism to retrieve information.

We also employ HERITRIX [15], one of major crawlers from open source project, for automatically collecting information from the Web. Given an initial URL address, the crawler will retrieve the corresponding Web page and extract all links found in the page. The process is then repeated for each found link until a termination condition is met.

In the proposed architecture, the search results are grouped using two-stage strategy. In the first stage, the system will attempt to hierarchically classify the search results based on a known classification category. If the category of some search results cannot be identified, then these results will be grouped by employing (unsupervised) clustering technique.

An ontology is utilized to help accurately classify the category of a Website. It contains the definition of information classes (categories) and their relationships (i.e., information structure). The definition of information class is used for Website classification. To speed up the hierarchical classification process, the identification of each Website category in the collection is performed in off-line mode. The classifier generates a metadata about the classification scheme for each Website (i.e., the Websites class and its path to the root in the ontology). The metadata provides the information of how a Website should be hierarchically classified.

Each document in the collection is also summarized in off-line mode. The document informative summary will be used for presentation of search results as an option to quickly view the content of search results before a user decides to download the corresponding document.

Given a search query, the search results returned by the LUCENE search engine are hierarchically organized. This hierarchical structure is then fed to an interface generator that allows users to navigate through the structure. The search results are presented to users as an indicative and informative summary. This will be discussed in later section.

4. Search Results Classification

One of the main features of the search engine architecture is hierarchically organizing search results according to their category. Because response time is critical, the efficiency of this process is of great importance. This paper tackles this issue by separating the hierarchical classification process into off-line and on-line modes.

The off-line classification is performed when a new document is downloaded by crawler. The document category is determined by calculating the similarity between document
and the description of concept category defined in the ontology. In this research, both the concept description (i.e., model of information class) in ontology and document descriptions are represented as vector space models. Given a feature vector of Web page $D$, the category of the document is determined as follows

$$c = \begin{cases} \arg \max \text{Sim}(s, D) \text{ if } \max \text{Sim}(s, D) \geq \theta \\ \text{unknown} \end{cases}$$

where

$$\text{Sim}(s, D) = \cosine(s, D) = \frac{s \cdot D}{\|s\| \|D\|}$$

and $S$ is a set of concept descriptions (models of information class) that are leaf nodes in the ontology. Hence, the topic of a Web page is the closest class whose similarity to the Webpage is greater or equal to a pre-determined threshold value (e.g., $\theta$). If the similarity of the closest class is below threshold, then Webpage category is declared to be unknown. If a Webpage category can be identified, the system will create a metadata associated with the classification results. The metadata stores information about the Webpage class and a sequence of its super classes leading to the root in ontology.

The on-line hierarchical classification is invoked when processing search results in responding to a query. A more detailed discussion of the hierarchical classification process is described in a separate paper [10].

On-the-fly clustering is applied for search results whose categories are unknown. There are many such clustering algorithms that have been developed. Research for determining the most suitable one for our purpose, particularly based on the work by Zheng et al [19], is in progress.

5. Search Results Presentation

The second research approach focuses on the development of summarization and interaction techniques that are specially designed to bring user searching experiences to the next level. The summarization technique will generate compact description of each search result, providing users with an accurate overview before deciding which Website to read.

This paper proposes two types of summary for viewing search results. The first type is indicative, query-relevance summary. It will be presented to users by default as snippet (inline summary) to indicate the document content. Indicative summary consists of two to three-line summaries that suggest the contents of original documents closely related to a user query. This summary facilitates a quick scanning among the search results. Informative summary is meant to represent the original document. It provides succinct description of the original document, providing an idea of what the whole content of document is all about. Informative summary can be designed into a manageable size and enables users to scan the document content.

The second summarization type is an informative, extractive and user oriented summary, providing users with more comprehensive, yet manageable document summary. Informative summary is generated through sentences selection from a document. Each sentence is scored based on the values of its features such as words and sentences location. The summary for each document is then generated from the top-scoring sentences, until a desired summary length is reached. Indicative summary is generated similarly, but is biased by user query. This summary is optional to users, and emerges as users hover over designated texts.

This research employs sentence extraction technique for generating both indicative and informative summaries. The summarization algorithm is as follows: (1) score each sentence, (2) sort the sentence based on score, (3) generate summary using $k$ highest scored sentence until a desired length is satisfied. Four aspects can be considered for sentence scoring: clustered of significant words, title term frequency, location and query biased. More detail summarization algorithm is described in a separate paper [18]. Figure 2 shows the snapshot of the search engine prototype.

6. Discussion

Ontology is at the heart of the classification process, specifically during the off-line classification mode. As pointed in the Introduction section, the use of ontology can improve the classification accuracy. However, it also has a drawback in that its applicability is limited to a specific domain. Therefore, the use of ontology, to be effective as intended, is only suitable for a domain-specific search engine such as the test bed that has been developed in this research.

Research for extending the applicability of domain ontology is in progress. In particular, this research is proposing ontology collection for extending the generality of search engine it supports. The ontology collection contains a large number of ontology covering a wide range of domains. There are at least two approaches that can be considered when processing a new query in this setting. In the first approach, the system will first have to retrieve $k$ most appropriate ontology based on a given query. Search results are then classified based on the retrieved ontology. This approach is efficient because the hierarchical classification process can focus on presumably the most appropriate ontology. It also, however, may miss considering the most appropriate concept defined in other (i.e., not retrieved) ontology. The second approach is to find, for each search result, the most appropriate concept category in every ontology in the collection. Contrary to the first approach, the second approach is less efficient. However, the second approach is more optimal in terms of classification accuracy because it considers every possible concept in the ontology collection.

The idea of using ontology collection as described above also presents another challenge. The ontology will have to be crafted manually by an expert. Creating a large number of ontology is too costly and cannot be possible to scale if it has to be handled alone. Fortunately, the idea of community collaboration such as those in Wikipedia and Open Source software communities can be applied to alleviate the scalability problem. In light of this, we are proposing community-based ontology development for supporting information classification. Community of experts can collaboratively contribute to develop ontology related to their
expertise. This will require a collaborative ontology creation environment, enabling a concern expert to initiate, develop, evaluate and collaborate in ontology creation project.

6. Concluding Remark

This paper has described a search engine architecture that is expected to be a prototype of next generation search engine. Due to a large number of relevant information that is available on the Web, a good web ranking algorithm is not enough. A search engine also needs a good presentation of search results, which allows its users to quickly scan and conveniently explore the search results. To achieve this objective, this research comes to a conclusion that an open, global-scale, community-based ontology for information classification is needed.

References

(20) http://www.google.com
(21) http://www.yahoo.com
(22) http://search.msn.com

Fig. 2 Snapshot of Search Engine Prototype
(23) http://www.vivisimo.com
(24) http://lucene.apache.org
(25) http://www.dmoz.org
(26) http://www.wikipedia.org